

## Model-Based Sieve Analysis

**Paul T. Edlefsen**

Fred Hutchinson Cancer Research Center, Seattle, WA 9809, U.S.A.

*email:* pedlefse@fhcrc.org

**SUMMARY:** We introduce a modeling framework and methods for Bayesian and frequentist assessment of site-specific sieving for analyzing genomic sequences of infecting pathogens in vaccine clinical trials.

**KEY WORDS:** categorical data analysis; sieve analysis; sieve effect; vaccine efficacy; Bayesian; breakthrough infection.

## 1. Introduction

Sieve analysis (Gilbert et al., 2001; Berman et al., 1997) aims to identify vaccine-induced effects on the distribution of infecting pathogens in participants of vaccine clinical trials by investigating the genomic sequences of infecting pathogens. In a properly conducted clinical trial, any observed genomic differences between vaccine-recipient infections and placebo-recipient infections are due either to chance alone or to the randomized treatment assignment. This article introduces statistical methods for evaluating and comparing these competing hypotheses.

The purpose of preventative vaccination is to elicit immune responses that target future infections. A successful vaccine induces immune responses that greatly reduce or even eliminate the ability of a pathogen to establish an infection or to thrive after infection. Since pathogens vary, vaccination against one is not always protective against another, and since some pathogens can vary greatly even within a single species, a vaccine may elicit protection against some but not all pathogens of any given type.

For simplicity of presentation, and because our analyses mainly address HIV and SIV viruses, we will use the term “virus” rather than “pathogen”, with the implicit understanding that the discussion applies equally to bacterial and fungal pathogens.

Many vaccines are designed to contain fragments of viral genomes (called the “vaccine insert” strains), with the intention that host immune responses against these fragments will cross-react with circulating viral strains. The degree of cross-reactivity is expected to correlate with protection: those circulating strains that are most similar to the vaccine inserts will be selectively filtered out (“sieved”) by the vaccine-induced immune response.

If the similarity of the genomic (RNA, DNA, or protein) sequence between the vaccine inserts and the circulating viruses is a good proxy for the relevant immunologic cross-reactivity, then viruses with sequences that are most similar to the insert sequences are

expected to be the most “filtered out” by the vaccine recipients’ immune systems. Sieve analysis seeks to quantify the sieve effect by directly comparing the genomic distributions of the viral sequences that infect vaccine recipients to those that infect placebo recipients.

In this article we focus on sieve analysis methods that seek to identify particular genomic loci exhibiting sieve effects. While global whole-gene or whole-genome comparisons are also valuable, discovery of an overall effect inevitably begs the question of which particular sequence features drive those effects. At a particular locus, each viral sequence contributes a single categorical datum, which might be a nucleotide or an amino acid, depending on the sequences being analyzed. Other categorical representations are possible, including physicochemical classes of amino acids (eg hydrophobicity), or structural properties of the genomic elements (eg being a part of an alpha-helix). While continuous-variable representations are also possible (such as physicochemical z-scales) , we focus on categorical representations, which are natural for these data. In particular, we will assume throughout this article that the viral sequences are aligned and translated, such that each discrete locus (or “site”) corresponds to a column of the multiple sequence alignment, with rows of the alignment corresponding single sequences, and each sequence at each locus has either one of the 20 possible amino acids or a gap indicator (for sequences having a deletion at the given locus). The assumption of amino acid data is made without loss of generality to other categorical representations, and the focus on individual sites does not preclude global testing based on agglomeration of individual locus-specific statistics.

Note that sieve effects are not the only explanation for observed differences between the viral sequences sampled from placebo recipients and vaccine recipients, since viruses continue to evolve after infection and may evolve differently in the presence of a vaccine-induced immune response. While some sieve analysis methods explicitly address the distinction between true sieving and post-infection differential evolution, we follow the example set

by Gilbert et al. (2008), to focus on evidence of a difference in the distributions of sequences in the two groups rather than on the cause of that difference. We assume that the sequences are collected sufficiently close to infection that any observed effects are true sieve effects, but acknowledge that the methods can also be used to identify post-acquisition differences if the sequences are collected at later time points. The differentiation of these two causes is beyond the scope of this paper.

Throughout, we assume that we have one sequence per subject and one vaccine insert sequence. Some vaccines are “multi-valent” in that they include multiple insert sequences; for this discussion we assume that such cases will be analyzed on a per-insert basis. It is not atypical to have multiple observations per subject, corresponding to multiple sequences of the pathogen. For the present discussion we will assume that a single sequence (per subject) has been chosen for analysis, either by choosing a subject-level consensus sequence or through some other means. Future work will address the potentially important contributions of rare-variant sequences and the simultaneous assessment of multiple vaccine insert strains.

One approach to identifying site-specific sieve effects is to apply standard survival analysis methods to infection timing data that is broken down by the genotype of the infecting strain, to determine differential genotype-specific vaccine efficacy (Prentice et al., 1978; Gilbert, 2000). This method can directly compare, for instance, vaccine efficacy against viruses that are insert-matched at a particular locus to efficacy against those that are a mismatch relative to the insert.

Other prototypical site-specific sieve analysis methods, introduced in Gilbert et al. (1998), apply maximum likelihood methods for standard categorical data models such as multinomial logistic regression and cumulative linear logit models.

The simplest method, introduced in Gilbert et al. (2008), compares the fraction of mismatches in the vaccine group to the fraction of mismatches in the placebo group. This test,

which compares the difference of these fractions to a null distribution estimated nonparametrically by permutation, can be conducted either on a per-site basis or by pooling some or all sites (to, for example, detect a sieve effect per viral gene). Variations of this test use Studentized statistics, and incorporate optional weights to allow use of sequence substitution matrices that reflect observed virus-specific amino acid composition and substitution frequencies. Others have applied this use of weights to reflect physico-chemical properties (such as amino acid hydrophobicity) and even to incorporate subject-specific weights based on subject genotype information with models predicting cytotoxic T-cell (CTL) binding affinities.

All of these methods consider univariate summary statistics and most resist distributional assumptions. While some approaches assume  $T$  distributions for estimates of distances or differences, most use nonparametric estimates of null distributions. This article is motivated by a desire to explicitly model the sieve effect phenomenon, with the goal of establishing a unifying framework for investigating sieve effects and for incorporating potentially complex covariate information. While we likewise embrace a healthy resistance to modeling assumptions, we argue that the phenomenon of sieving implies certain distributional constraints which can be modeled.

Consider the simple mismatch-fraction statistic of Gilbert et al. (2008), for a single site. From a model-based perspective, this approach considers a null distribution in which each subject's amino acid differs from the insert amino acid with a probability  $p$  that does not depend on the subject's vaccine treatment assignment. That is, it dichotomizes the amino acids into either a mismatch, 1, or a match, 0, and considers each subject's match/mismatch status as an independent random variable with a  $\text{bernoulli}(p)$  distribution. Under the alternative, the model is the same except that the probability  $p_z$  of a mismatch depends on the subject's treatment assignment  $z$ , which we can arbitrarily designate as  $z = 0$  for placebo

recipients and  $z = 1$  for vaccine recipients. Then any standard test of  $p_1 = p_0$  vs  $p_1 \neq p_0$  can be used to test the null hypothesis, and a sieve effect will be identified if  $(p_1 - p_0)$  is both statistically significantly different from 0 and of sufficient magnitude to be biologically relevant.

Next consider the weighted variant, also introduced in Gilbert et al. (2008), in which each subject-specific weights replace match/mismatch indicators. These weights are grouped by treatment assignment, and the Studentized difference is used as a test statistic (with a 0 difference under the null distribution). In the special case in which these weights are substitution frequencies, reflecting the probability that the insert residue will be substituted by the subject's observed residue, this approach models a sieve effect as differential selective pressure induced by the vaccine.

In this article we introduce a framework and a set of models for a model-based approach to sieve analysis. The aim is to ground categorical sieve analysis methods in a statistical modeling framework that corresponds to the scientific understanding of the phenomenon, and to use this framework to expand on previously established methods. The goals include both hypothesis testing (eg. Is there a sieve effect?) and parameter estimation (eg. How strong is the sieve effect?). These two goals may be unified by, for instance, rejecting the hypothesis of no sieve effect whenever the estimate of the strength of the sieve effect confidently excludes zero.

In addition to providing a formalism for evaluating and extending models of sieving, the methods and framework that we introduce in this article contribute to the practice of sieve analysis by incorporating features that were, to our knowledge, previously unavailable. We introduce Bayesian and frequentist-Bayesian hybrid methods that allow for direct assessment of posterior probabilities of sieving (and other features), and that accommodate incorporation of prior information such as the distribution of amino acids among circulating viral strains,

and subject-specific covariates such as vaccine immunogenicity measurements. We argue and demonstrate through simulation studies that these methods have greater power to detect sieve effects than those that ignore this information, and that explicit modeling of sieve effects enables more direct interrogation of the nature of the effect (via, for instance, the posterior probability that a sieve effect is of the “insert-only” variety, as discussed below).

As an initial illustration of the model-based perspective, let us now reconsider the mismatch-fraction statistic. Suppose that we wish to explicitly represent the expectation that the fraction of mismatches should increase in the vaccine group in the presence of a sieve effect. This could be accommodated by conducting one-sided versions of the (non-model-based) tests, but in a model-based framework we can instead explicitly model the sieve effect as the amount by which the probability of a match is reduced due to vaccination (this is what we will call the “sieve effect strength”  $p_s$ ). This model amounts to a reparameterization of the above representation of the match-mismatch test, such that under the alternative model,  $p_1 = p_0(1 - p_s)$ .

This simple model can be extended by considering the distribution of non-insert amino acids. Since we understand sieve effects as resulting from vaccine-induced selective filtering of potentially-infecting viruses based on their similarity to the vaccine insert, we can interpret the dichotomous model as defining similarity based on amino acid identity at the considered locus, with no interest in the non-insert distribution. Taking the model of selective filtering literally, we might expect that the distribution of non-insert amino acids in infecting sequences, conditional on being a non-insert amino acid, is no different from that of the circulating strains (since by this logic, filtering depends on the insert/non-insert dichotomy only). If we believe this “insert-only” model of sieving, then we might do better to model it explicitly by representing our expectation that the non-insert distribution is the same in both treatment groups.

Perhaps instead we argue that the reason to dichotomize is that we wish to remain agnostic about the non-insert distribution in the alternative model, because the immune response that successfully filters out insert-matched viruses may target insert-mismatched viruses too, perhaps based on a physico-chemical relatedness to the insert amino acid. This case, too, can be explicitly modeled by including the non-insert amino acids, with a distribution under the alternative that differs from that under the null. This alternative non-insert distribution could be specified as, for instance, a substitution matrix reflecting physico-chemical relatedness, or it could be modeled as an unknown distribution (following eg a Dirichlet prior).

If we consider that some vaccine recipients become infected despite their successful immune targeting of insert-like viruses, while others simply are not able to mount such a response (due to an imperfect vaccine), we might expect that some fraction of the vaccinated subjects will be effectively placebo-like. The sieve effect strength parameter ( $p_s$ ) is then interpretable as the fraction of subjects who have infecting viral distributions affected by a sieve effect. Then the non-insert distribution becomes a mixture: the non-sieved part, which looks like the placebo distribution, and the sieved part, which potentially looks different.

The framework is easily extended to incorporate other evidence about subject-specific responses to the vaccine. By definition, placebo recipients are incapable of having sieve effects. It may also be the case that some of the vaccine recipients are incapable of sieving (those with no “take” of the vaccine, for instance). If we have information suggesting which of the vaccine recipients are capable of sieving at a particular locus (immunogenicity data after vaccination, for instance), then we might further introduce explicit indicators of subjects’ ability-to-sieve. These *sieveability* indicators, which are in general not observed, reflect a potential for sieving rather than actual sieving. Thus we conceptually separate *sieveability* from actually being “sieved”. Even when all vaccine recipients have the ability to sieve, a



sieve effect strength can be less than one, indicating that some of the viruses that match the insert are still infecting sievable vaccine recipients.

The overall approach can be summarized as a latent-variable framework for modeling the distribution of amino acids of viral sequences. The primary latent variables are called *sieved*. These form a matrix of per-subject, per-locus indicators of having been affected by a sieve effect. At each locus, these indicators are modeled as independent bernoulli( $p_s$ ) random variables for “sievable” subjects, and are 0 otherwise. The model assumes conditional independence among the amino acids at each locus for each subject, given these latent *sieved* indicators.

The model framework assumes that the distribution of amino acids at a particular locus (when not *sieved*) is common across all not-sieved subjects, including the placebo subjects (who are by design not vaccinated and by definition not sieved). It also assumes that the amino acid distribution at any given locus is common across sieved subjects. Future models could relax this assumption by accounting for the clade of the infecting strains or for subject-specific covariates such as HLA alleles, either of which could alter the distribution of observed sequences even in the absence of a sieve effect.

This framework connects covariate information to sequence data solely through the optional additional latent variables indicating per-subject, per-locus *sieveability*. In the sieve analyses of the HVTN 502 (STEP) and HVTN 503 (Phambili) HIV-1 vaccine clinical trials, we use the HLA covariates to predict t-cell epitopes, and the confidence values associated with these epitopes can be used directly for prior probabilities of the *sieveability* indicators (by converting each confidence value to a probability and then calculating the per-locus, per-subject probability that at least one t-cell epitope is real). If we additionally compute binding energy changes associated with HLA-specific CTL escape mutations, these can be incorporated into the *sieveability* indicator probabilities (which is effectively equivalent to

allowing different amino acid distributions across the sieved subjects at a locus, since the effect would depend on the subject’s observed amino acid(s) at the locus). For vaccines in which humoral immune responses are expected to play a role, such as in the RV144 HIV-1 vaccine clinical trial (the “Thai Trial”), various additional considerations could be used to affect the *sieveability* probabilities, including locus-specific factors reflecting known and predicted antibody binding regions, as well as subject-and-locus-specific factors estimated using targeted immunological assays (such as the gp70-V1V2 assay, which measures subject-specific antibody binding to a particular region of the HIV-1 envelope protein). Arbitrarily complex models for the *sieveability* indicator probabilities are possible, including generalized linear models (future work). The framework allows for methodological separation of the covariate-dependent aspects of the model from the covariate-independent aspects. This is useful, since the former are likely to differ across trials and with time.

In addition to the general framework, we introduce a likelihood-ratio testing framework with a shared null model and three specific alternative models, each of which comes in two major flavors, depending on how placebo-recipient data is used. Each of these models can be used in a full Bayesian analysis or in a frequentist analysis, and each can accommodate a hierarchical perspective of the null-distribution category probabilities, or alternatively a “non-hierarchical” perspective. All can accommodate arbitrary probabilities of *sieveability* (but each is simpler with the assumption that all vaccine recipients are sieveable at the given locus). The alternative models differ only in their treatment of the conditional distribution among the non-insert categories: the “insert-only” model assumes that this conditional distribution is identical to that of the null model, whereas the “noninsert-unconstrained” model assumes an arbitrary (other) distribution. The “noninsert-monotonic” model is just like the “noninsert-unconstrained” model, except that non-sieved subjects follow the null model’s conditional non-insert distribution (that is, the reallocation of mass among the non-

insert categories applies only to *sieved* subjects). These alternative models can be combined into a single analysis by introducing prior probabilities over the alternative models; for instance we typically use the “insert-only” and “noninsert-monotonic” models together, with interest in the posterior probability of “insert-only”ness. The model flavors are “one-phase”, which use all of the data in one analysis, and “two-phase”, which first use the placebo-recipient data to update the prior null-model distribution and then explicitly evaluate only the vaccine-recipient data. These models, flavors, and analysis approaches are described in further detail below.

## 2. Mathematical details

In this section we use the following notational conventions: emboldened variables (such as  $\mathbf{Y}$ ) represent vectors; non-bold variables (such as  $Y_j$ ) represent scalars. Probability mass functions for the binomial and multinomial distributions are represented as  $P_b(\cdot)$  and  $P_m(\cdot)$ , respectively. Probability density functions for the beta and Dirichlet distributions are represented as  $dF_\beta(\cdot)$  and  $dF_d(\cdot)$ , respectively. Probability parameters are represented by the lowercase values  $p$  and  $q$  (or  $\mathbf{p}$ , for probability vectors defined on a simplex) with additional subscripts for further differentiation (such as  $p_s$  versus  $p_r$ ). Parameters of prior distributions are represented by the greek letters  $\alpha$  and  $\gamma$ . A superscripted “ $N$ ” on a vector (such as  $\alpha_0^N$ ) denotes the subvector that excludes element 1 (the insert element), and a superscripted “ $N$ ” on a scalar (such as  $\alpha_0^N$ ) denotes the sum of the non-insert elements of the corresponding vector (such that  $\alpha_0^N = \sum_i \alpha_{0,i}^N$ ). Other notation is defined where it is first used.

In this analysis we consider only a single column (site) of multiply-aligned sequences. The data at this site are of the form  $Y_j \in 1, \dots, k$  where  $k$  is the number of categories (typically in our analyses there are 21 categories, corresponding to the 20 amino acids and to the gap indicator; for nucleotide-level analyses we allow 5 categories to represent the four nucleotides

and the gap indicator), and  $j \in 1, \dots, n$  identifies the subject. We will assume that each subject has exactly one value  $Y_j$ .

Our analysis, be it frequentist or Bayesian, requires computing likelihoods of the data under null and alternative models  $M_0$  and  $M_a$ . In the frequentist analysis we compare the ratio of likelihoods or, equivalently, the difference in log-likelihoods. When being Bayesian, we compute the posterior probability of the alternative model as proportional to the likelihoods (weighed by priors), since  $P(M_a|\mathbf{Y}) = \frac{P(\mathbf{Y}|M_a)P(M_a)}{P(\mathbf{Y}|M_0)P(M_0)+P(\mathbf{Y}|M_a)P(M_a)}$ . In either case we begin by computing the likelihoods  $P(\mathbf{Y}|M_a)$  and  $P(\mathbf{Y}|M_0)$ .

Under the null, we model all of the data as independently drawn from a multinomial distribution:  $(\mathbf{Y}|M_0) \sim \text{multinomial}(\mathbf{p}_0)$ , with category probabilities  $\mathbf{p}_0 = (p_{0,i} : i \in 1, \dots, k)$  either given as a pre-specified model parameter or averaged over a Dirichlet prior  $\mathbf{p}_0 \sim \text{Dirichlet}(\boldsymbol{\alpha}_0)$ , with a given hyperparameter  $\boldsymbol{\alpha}_0 \in \mathcal{R}^k$ . In the latter case, the null likelihood is given by  $P(\mathbf{Y}|M_0, \boldsymbol{\alpha}_0) = \int_{\mathbf{p}_0} P_m(\mathbf{Y}|\mathbf{p}_0) dF_d(\mathbf{p}_0|\boldsymbol{\alpha}_0)$ . Thus in this case  $\mathbf{Y}$  has a Dirichlet-multinomial distribution:  $(\mathbf{Y}|M_0, \boldsymbol{\alpha}_0) \sim \text{Dirichlet-multinomial}(\boldsymbol{\alpha}_0)$ .

Let us now consider the vaccination status of each subject. It will become clear that if we're being frequentist, the contributions to the null and alternative likelihoods from the placebo recipients are identical, so placebo recipients are effectively ignored in the analysis. As discussed above, in this article we investigate both the “one-phase” case in which we wish to directly model all of the data, and also the “two-phase” case, in which we instead use the placebo-recipient data to estimate  $\boldsymbol{\alpha}_0$  (or  $\mathbf{p}_0$ ) and explicitly perform our comparison of models using only the vaccine-recipient data. In the two-phase case, we can make use of the placebo data by using the posterior distribution of  $\boldsymbol{\alpha}_0$  given the placebo-recipient values. If we order the subjects such that those with identifiers  $i \in 1, \dots, n_v$  are vaccine recipients and those with identifiers in  $(n_v + 1), \dots, n$  are placebo recipients (where  $n$  is  $n_v + n_p$ ), then we may estimate  $(\boldsymbol{\alpha}_0|\mathbf{Y}_p = Y_{(n_v+1), \dots, n})$  using the  $n_p$  placebo recipients' data via the

typical conjugate Bayesian posterior  $(\boldsymbol{\alpha}_0 | \mathbf{Y}_p, \boldsymbol{\gamma}) \sim \text{Dirichlet}(K(\mathbf{Y}_p) + \boldsymbol{\gamma})$ , where  $\boldsymbol{\gamma} \in \mathcal{R}^k$  are the hyperparameters serving as priors for  $\boldsymbol{\alpha}_0$  (before observing the placebo data  $\mathbf{Y}_p$ ), and where  $K(\cdot) : \mathcal{I}^m \mapsto \mathcal{I}^k$  is the function that categorizes a vector of  $m$  integers  $\in \{1, \dots, k\}$  into counts for each of the  $k$  categories. If we prefer to directly specify  $\mathbf{p}_0$ , then we can use the maximum likelihood estimator (MLE), which is  $\hat{\mathbf{p}}_0 \propto K(\mathbf{Y}_p)$ , or the posterior mean of  $(\boldsymbol{\alpha}_0 | \mathbf{Y}_p, \boldsymbol{\gamma})$ , which is  $\tilde{\mathbf{p}}_0 \propto K(\mathbf{Y}_p) + \boldsymbol{\gamma}$ .

Note that the decision to treat  $\mathbf{p}_0$  as known or to model it as  $(\mathbf{p}_0 | \boldsymbol{\alpha}_0) \sim \text{Dirichlet}(\boldsymbol{\alpha}_0)$  isn't simply a frequentist/Bayesian distinction. Using  $\boldsymbol{\alpha}_0$  in a frequentist framework is effectively treating  $\mathbf{p}_0$  as a random effect and integrating it out as a nuisance parameter. On the other hand, using the posterior mean estimate for  $\mathbf{p}_0$  is a form of Empirical Bayes (one that is not subject to the oft-cited concern of multiple “uses” of data, since the placebo data used to estimate  $\mathbf{p}_0$  are not evaluated in the model). The estimator  $\tilde{\mathbf{p}}_0$  has a frequentist interpretation if the values  $\boldsymbol{\gamma}$  are treated as “pseudocounts”, which here are advisable since  $\hat{\mathbf{p}}_0$  will assign zero probability to any category not observed in the placebo data. In practice, since pathogen sequences are evolutionarily conserved, in many cases all but one or two of the amino acid categories are rarely observed, and as our sieve analyses are generally conducted with far-from-asymptotic quantities of data, unobserved categories are typical. As discussed below, it may be possible to collect additional data on circulating pathogens to augment estimation of  $\mathbf{p}_0$ , but this will often still require use of pseudocounts to handle unobserved categories.

We now proceed to the likelihood of the (vaccine recipient) data under the alternative model  $M_a$ . We model a sieve effect as a reduction in the probability of the insert category and a redistribution of that mass among the other categories. The strength  $p_s$  of the sieve effect is a parameter of interest, which we must either estimate or integrate-over. Without loss of generality, assume that the categories are ordered such that category 1 is the insert category.

Then the probability of the insert category under the null is  $p_{0,1}$  and under the alternative (given  $p_s$ ) it is  $p_{0,1}(1 - p_s)$ . Of course the probabilities must sum to one; we redistribute that removed mass of  $p_{0,1}p_s$  either proportionally to the corresponding elements of  $\mathbf{p}_0$ , or proportional to an alternative probability vector  $\mathbf{p}_a$ . These options reflect a “insert-only” and an “noninsert-monotonic” model of sieving, respectively. Note that the insert-only approach is equivalent to using  $\mathbf{p}_a = \mathbf{p}_0$  in the “noninsert-monotonic” model when  $\mathbf{p}_0$  is considered as a fixed quantity.

### 2.1 The insert-only alternative model $M_a^{io}$

Ironically, this dependence of  $\mathbf{p}_a$  on the values of  $\mathbf{p}_0$  results in a relatively simple calculation for the insert-only alternative likelihood, even in the scenario in which we model  $\mathbf{p}_0$  as a random effect. Thus we begin with this case before proceeding to the case in which  $\mathbf{p}_a$  is given (and does not depend on  $\mathbf{p}_0$ ). For simplicity we will not consider other hierarchical models for  $\mathbf{p}_a$ , though they are potentially interesting and could be addressed by further extensions to the model.

The reason for the relative simplicity of the insert-only alternative model is that the conditional probabilities among the non-insert categories do not depend on  $p_s$  and are the same under the null and alternative models. Thus in this case the likelihood ratio and the posterior probability of the alternative model depend only on the number of subjects in each treatment group that match the insert and the number that do not match. This analysis effectively dichotomizes the data. If we denote by  $\alpha_0^N$  the sum of the non-insert values of  $\boldsymbol{\alpha}_0$ , so that  $\alpha_0^N = \sum_{i=2} \alpha_{0,i}$ , then in this insert-only alternative scenario the null-model likelihood (for a hierarchical approach to  $\mathbf{p}_0$ ) reduces to

$$P(\mathbf{Y}|M_0, \boldsymbol{\alpha}_0) = \int_{\mathbf{p}_{0,1}} P_b(K_I(Y)|\mathbf{p}_0) dF_{\beta}(p_{0,1}|\alpha_{0,1}, \alpha_0^N),$$

where the function  $K_I(\cdot) : \mathcal{I}^m \mapsto \mathcal{I}^2$  counts the number of insert and non-insert values in a

vector of counts. The alternative-model likelihood for a fixed sieve effect strength  $p_s$  is

$$P(\mathbf{Y}|M_a^{io}, \boldsymbol{\alpha}_0, p_s) = \int_{p_{0,1}} P_b(K_I(\mathbf{Y}_p)|p_{0,1})P_b(K_I(\mathbf{Y}_v)|p_{0,1}(1-p_s))dF_\beta(p_{0,1}|\alpha_{0,1}, \alpha_0^N),$$

where the placebo data  $\mathbf{Y}_p$  are modeled as under the null, and the vaccine data  $\mathbf{Y}_v$  have reduced probabilities assigned to the insert category. Integrating  $p_s$  over an arbitrary beta( $\boldsymbol{\alpha}_s$ ) prior yields

$$P(\mathbf{Y}|M_a^{io}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_s) = \int_{p_s} P(\mathbf{Y}|M_a^{io}, \boldsymbol{\alpha}_0, p_s)dF_\beta(p_s|\boldsymbol{\alpha}_s).$$

In principle the values  $\boldsymbol{\alpha}_s$  could be set to reflect prior evidence about the sieve effect strength, but care must be taken with regard to identifiability between the distribution of  $p_s$  and the probability  $P(M_a)$  of the alternative model. Taken together, these effectively yield a zero-inflated prior for  $p_s$ , so non-negligible prior mass on  $p_s$  near 0 might yield difficult-to-interpret posteriors. We typically use a uniform prior ( $\boldsymbol{\alpha}_s = (1, 1)$ ), which leads to a nuisance-parameter interpretation of  $p_s$  in the frequentist context.

If we are treating  $\mathbf{p}_0$  as fixed and known, for example by using  $\mathbf{p}_0 \propto \boldsymbol{\alpha}_0$ , then the placebo parts of the above equations cancel and the analysis ignores the placebo data.

## 2.2 The noninsert-monotonic alternative model $M_a^{nm}$

The simplicity of the insert-only model is due to its effective ignorance of the distribution of non-insert categories. If we believe that the distribution of non-insert categories differs in the presence of a sieve effect, then attention to the non-insert categories should yield more effective models and more powerful tests. It may be tempting to assume that under the alternative model, the non-insert distribution is arbitrarily different than the null model's multinomial (or Dirichlet-multinomial) distribution, and we present such an approach for comparison, below ( $M_a^{nu}$ ). Here, however, we restrict the model to reflect our expectation that subjects for whom the vaccine failed to induce a sieve effect will have infecting viruses

with a distribution that matches the null distribution. That is, since only the subjects for whom there is a sieve effect will have a different distribution of infecting viruses, only the redistributed mass  $(p_{0,1}p_s)$  will be differently-distributed under the alternative model. This ensures that if  $p_s = 0$ , the alternative model is identical to the null model (which is of course also true for  $M_a^{io}$ , but as we will see, it is not so for  $M_a^{nu}$ ). An implication of this constraint is that the marginal probability for any non-insert category  $i$  under the alternative model ( $p_{a,i}$ ) is at least as large as its marginal probability under the null (this is the monotonicity constraint that  $p_{a,i} \geq p_{0,i} \forall i > 1$ ).

Note that the insert-only alternative model described above need not be concerned with the conditional distribution of non-insert categories, since the redistributed mass is assumed to be distributed according to the same conditional probabilities as the null. For clarity, we here repeat the presentation of the insert-only alternative likelihood with the non-insert categories included. Letting  $\alpha_0^N = (\alpha_{0,i} : i \in 2, \dots, k)$  be the non-insert values in  $\alpha_0$ , and similarly letting  $\mathbf{p}_0^N = (p_{0,i} : i \in 2, \dots, k)$  be the non-insert values in  $\mathbf{p}_0$ , we have

$$\begin{aligned}
P(\mathbf{Y} | M_a^{io}, \alpha_0, p_s) = & \\
& \int_{\mathbf{p}_0^N} P_m(K_N(\mathbf{Y}) | \mathbf{p}_0^N) dF_d(\mathbf{p}_0^N | \alpha_0^N) \\
& \times \int_{p_{0,1}} P_b(K_I(\mathbf{Y}_p) | p_{0,1}) P_b(K_I(\mathbf{Y}_v) | p_{0,1}(1 - p_s)) dF_\beta(p_{0,1} | \alpha_{0,1}, \alpha_0^N),
\end{aligned} \tag{1}$$

where the function  $K_N(\cdot) : \mathcal{I}^m \mapsto \mathcal{I}^{k-1}$  counts the values in a vector of counts, but ignores values of 1 (the insert category). In essence, we are treating the multinomial model over all counts as the product of two models: a binomial to determine whether the category is the insert category, and a (smaller) multinomial model for the conditional distribution over the remaining categories.

Since the non-insert part of (1) does not depend on  $p_s$ , it factors out when we integrate



over  $p_s$ , and cancels when we consider the null model likelihood as

$$\begin{aligned}
 P(\mathbf{Y}|M_0, \boldsymbol{\alpha}_0) = & \\
 & \int_{\mathbf{p}_0^N} P_m(K_N(\mathbf{Y})|\mathbf{p}_0^N) dF_d(\mathbf{p}_0^N|\boldsymbol{\alpha}_0^N) \\
 & \times \int_{p_{0,1}} P_b(K_I(\mathbf{Y})|p_{0,1}) dF_\beta(p_{0,1}|\alpha_{0,1}, \alpha_0^N).
 \end{aligned} \tag{2}$$

In the noninsert-monotonic-alternative likelihood, the non-insert part depends on both  $p_s$  and  $p_{0,1}$ , since the non-insert probability vector  $\mathbf{q}_a^N$ , which for convenience we'll index from 2 to  $k$ , is  $\mathbf{p}_0^N + p_{0,1}p_s\mathbf{p}_a^N$  (where again we use the notation  $\mathbf{p}_a^N = (p_{a,i} : i \in 2, \dots, k)$ , and assume or impose the restriction that  $\sum_{i=2}^k p_{a,i} = 1$ , so that  $\sum_{i=2}^k q_{a,i} = 1 - p_{0,1}(1 - p_s)$ ). Thus we get

$$\begin{aligned}
 P(\mathbf{Y}|M_a^{nm}, \boldsymbol{\alpha}_0, \mathbf{p}_a, p_s) = & \\
 & \int_{p_{0,1}} \left( \begin{aligned} & P_b(K_I(\mathbf{Y}_p)|p_{0,1})P_b(K_I(\mathbf{Y}_v)|p_{0,1}(1 - p_s)) \\ & \times \int_{\mathbf{p}_0^N} (P_m(K_N(\mathbf{Y}_p)|\mathbf{p}_0^N)P_m(K_N(\mathbf{Y}_v)|\mathbf{q}_a^N) dF_d(\mathbf{p}_0^N|\boldsymbol{\alpha}_0^N)) \end{aligned} \right) dF_\beta(p_{0,1}|\alpha_{0,1}, \alpha_0^N),
 \end{aligned} \tag{3}$$

which in the “non-hierarchical”  $\mathbf{p}_0$  scenario simplifies to

$$\begin{aligned}
 P(\mathbf{Y}|M_a^{nm}, \mathbf{p}_0, \mathbf{p}_a, p_s) = & \\
 & P_b(K_I(\mathbf{Y}_p)|p_{0,1})P_b(K_I(\mathbf{Y}_v)|p_{0,1}(1 - p_s)) \times \\
 & P_m(K_N(\mathbf{Y}_p)|\mathbf{p}_0^N)P_m(K_N(\mathbf{Y}_v)|\mathbf{q}_a^N).
 \end{aligned} \tag{4}$$

### 2.3 The noninsert-unconstrained alternative model $M_a^{nu}$

As promised, we now investigate a different approach to the non-insert-only reallocation alternative model, in which the non-insert distribution in the alternative model is given by a multinomial with probabilities  $\mathbf{p}_a^N$ . Unlike the insert-only and noninsert-monotonic models discussed above, in this scenario a “sieve effect” can be identified in data in which the observed frequency of the insert category is identical between placebo and vaccine groups, since even if the sieve effect strength is  $p_s = 0$ , the alternative differs from the null. We call

this the “noninsert-unconstrained” model because the entire non-insert mass is reallocated, rather than just the sieved mass as in the noninsert-monotonic model. Note that if  $\mathbf{p}_a = \mathbf{p}_0$ , then (when  $\mathbf{p}_0$  is fixed and known) all three of these models coincide.

The likelihood of the data given the noninsert-unconstrained alternative model is given by

$$\begin{aligned}
 P(\mathbf{Y}|M_a^{nu}, \boldsymbol{\alpha}_0, \mathbf{p}_a, p_s) = & \\
 & P_m(K_N(\mathbf{Y}_v)|\mathbf{p}_a^N) \int_{\mathbf{p}_0^N} P_m(K_N(\mathbf{Y}_p)|\boldsymbol{\alpha}_0^N) dF_d(\mathbf{p}_0^N|\boldsymbol{\alpha}_0^N) \\
 & \times \int_{p_{0,1}} P_b(K_I(\mathbf{Y}_p)|p_{0,1}) P_b(K_I(\mathbf{Y}_v)|p_{0,1}(1 - p_s)) dF_\beta(p_{0,1}|\alpha_{0,1}, \alpha_0^N).
 \end{aligned} \tag{5}$$

#### 2.4 Latent indicators of sievability

The logic that differentiates the noninsert-monotonic alternative from the noninsert-unconstrained alternative hinges on the notion that some subjects will be capable of sieving at a location and others will not. Those who are not capable of sieving are expected to have null-like distributions, just as if they had received the placebo. This notion is rooted in the principle that a vaccine-induced humoral (antibody) or cell-mediated (t-cell) response is a prerequisite for sieving, and that not all subjects are equally responsive. In HIV-1 vaccine trials, for example, the induction of binding antibodies has varied widely across individuals, as has the induction of cd8+ t-cell responses. Cell-mediated responses are known to depend on a subject’s genotype (HLA type), and can be predicted with some accuracy for some specific peptide fragments. The STEP HIV-1 vaccine trial induced no humoral responses, so its effects are believed to be entirely t-cell-mediated. For sieve analysis of the STEP trial data, this might suggest excluding vaccine subjects who lack the appropriate HLA type to generate an immune response at a given site.

In this section we introduce explicit notation for indicators of sievability. Since even subjects capable of sieving are not guaranteed to sieve, it is important to stress that these indicators are meant to exclude incapable subjects, not to indicate which subjects have sieve

effects. We treat these indicators as unknown (latent) values. The idea, which we will see is borne out by the simulation trials, is that if you can provide reasonably-accurate probabilities for these latent indicators, the power to identify sieve effects is greatly improved.

We now revisit and extend the three alternative models (insert-only, noninsert-monotonic, and noninsert-unconstrained) to accomodate subject-specific latent indicators  $\mathbf{l} = (l_i : i \in 1, \dots, n)$  with corresponding probabilities  $\mathbf{p}_r$  such that  $P(l_i = 1) = p_{r,i}$ . By convention we will allow these latent variables for even the placebo subjects, but set  $p_{r,i} = 0 \forall i \in (n_v + 1), \dots, n$ . We will not presently concern ourselves with how the vaccine-subject indicator probabilities are set, but in the results section we provide examples using HLA prediction and antibody measurements.

Before we proceed, we must address the notion of exchangeability. In our earlier expression of the null model as a multinomial (or Dirichlet-multinomial), we were implicitly treating the data as count data, though the multinomial coefficient that accomplishes this is constant and cancels across models, so it would be equivalent to treat the data as non-exchangeable. In that case the null likelihood could be represented by a product over subjects:

$$P(\mathbf{Y}|M_0, \boldsymbol{\alpha}_0) = \int \left( \prod_i p_{0,i} \right) dF_d(\mathbf{p}_0|\boldsymbol{\alpha}_0),$$

which can be calculated as the Dirichlet-multinomial probability divided by the multinomial coefficient. Using a non-exchangeable representation will free us to model the subject-level data directly without concern for the many ways in which (for example) 30 of the 44 vaccine recipients' sequences could match the insert category.

The insert-only alternative model likelihood, averaging over the latent sievable indicators,

is

$$\begin{aligned}
P(\mathbf{Y}|M_a^{io}, \boldsymbol{\alpha}_0, \mathbf{p}_r, p_s) = & \int_{\mathbf{p}_0^N} \frac{1}{C(K_N(\mathbf{Y}))} P_m(K_N(\mathbf{Y})|\mathbf{p}_0^N) dF_d(\mathbf{p}_0^N|\boldsymbol{\alpha}_0^N) \\
& \times \int_{p_{0,1}} \left( \prod_{i:Y_j=1} \left( p_{r,i}p_{0,1}p_s + (1-p_{r,i})p_{0,1} \right) \right. \\
& \left. \times \prod_{i:Y_j \neq 1} \left( p_{r,i}(1-p_{0,1}p_s) + (1-p_{r,i})(1-p_{0,1}) \right) \right) dF_\beta(p_{0,1}|\alpha_{0,1}, \alpha_0^N),
\end{aligned} \tag{6}$$

where we use the notation  $C(\cdot)$  to represent the multinomial coefficient. When we treat  $\mathbf{p}_0$  as a “fixed-effect”, this simplifies to

$$\begin{aligned}
P(\mathbf{Y}|M_a^{io}, \mathbf{p}_0, \mathbf{p}_r, p_s) = & \frac{1}{C(K_N(\mathbf{Y}))} P_m(K_N(\mathbf{Y})|\mathbf{p}_0^N) \\
& \times \prod_{i:Y_j=1} \left( p_{r,i}p_{0,1}p_s + (1-p_{r,i})p_{0,1} \right) \prod_{i:Y_j \neq 1} \left( p_{r,i}(1-p_{0,1}p_s) + (1-p_{r,i})(1-p_{0,1}) \right).
\end{aligned} \tag{7}$$

Note that as before, the non-insert parts of (6) and (7) factor, since they do not depend on  $p_s$  or on the latent sievable indicators  $\mathbf{l}$ .

The noninsert-monotonic model likelihood, averaging over latent sievable indicators, is given by

$$\begin{aligned}
P(\mathbf{Y}|M_a^{nm}, \boldsymbol{\alpha}_0, \mathbf{p}_a, \mathbf{p}_r, p_s) = & \int_{\mathbf{p}_0} \prod_{i:Y_j=1} \left( p_{r,i}p_{0,1}p_s + (1-p_{r,i})p_{0,1} \right) \prod_{i:Y_j \neq 1} \left( p_{r,i}q_{a,Y_j} + (1-p_{r,i})p_{0,Y_j} \right) dF_d(\mathbf{p}_0|\boldsymbol{\alpha}_0),
\end{aligned} \tag{8}$$

where again  $q_{a,i} : i \in 2, \dots, k$  are the non-insert category probabilities for subjects who can sieve at this site.

The noninsert-unconstrained alternative model likelihood, averaging over the latent siev-

able indicators, is

$$\begin{aligned}
P(\mathbf{Y}|M_a^{nu}, \boldsymbol{\alpha}_0, \mathbf{p}_r, p_s) = & \\
& \frac{1}{C(K_N(\mathbf{Y}_v))} P_m(K_N(\mathbf{Y}_v)|\mathbf{p}_a^N) \\
& \times \int_{\mathbf{p}_0^N} \frac{1}{C(K_N(\mathbf{Y}_p))} P_m(K_N(\mathbf{Y}_p)|\mathbf{p}_0^N) dF_d(\mathbf{p}_0^N|\boldsymbol{\alpha}_0^N) \\
& \times \int_{p_{0,1}} \left( \prod_{i:Y_j=1} \left( p_{r,i} p_{0,1} p_s + (1 - p_{r,i}) p_{0,1} \right) \right. \\
& \quad \left. \times \prod_{i:Y_j \neq 1} \left( p_{r,i} (1 - p_{0,1} p_s) + (1 - p_{r,i}) (1 - p_{0,1}) \right) \right) dF_\beta(p_{0,1}|\alpha_{0,1}, \alpha_0^N).
\end{aligned} \tag{9}$$

### 3. Implementation and Power

Our implementation of these methods, in the R statistical programming language, supports all permutations of the three alternative models (“insert-only”, “noninsert-monotonic”, and “noninsert-unconstrained”) with either flavor (“one-phase” or “two-phase”) and either option for  $\mathbf{p}_0$  (“non-hierarchical” or “hierarchical”), except that we haven’t implemented the combination of “hierarchical” and “noninsert-monotonic” (both because of the relative difficulty of implementation due to the requirement of integrating over all 19 free parameters of the conditional non-insert distribution  $\mathbf{p}_0^N$ , and because our simulation study investigations suggest little added benefit of the hierarchical treatment vs the “non-hierarchical” treatment for the “insert-only” and “noninsert-unconstrained” models). This code will be made available as part of a new package (“sieve”) on CRAN, which will also include our implementation of the Gilbert et al. (2008) method.

In a frequentist setting, we use likelihood ratio tests to compare  $M_0$  and a particular alternative model  $M_a$ . By Wilks’ theorem, if we use the MLE of  $p_s$  in evaluating the alternative model, then under the null,  $-2$  times the log-likelihood ratio comparing  $M_0$  to  $M_a^{io}$  (or  $M_a^{nm}$ ) is asymptotically chi-squared-distributed with 1 degree of freedom ( $p_s$  is the one parameter restricted by  $M_0$ ;  $\mathbf{p}_a$  can be considered a parameter of the null model, though

it is never used because  $p_s = 0$  there). Note that  $M_0$  is not nested within  $M_a^{nu}$ , since even when  $p_s = 0$ , the models differ, so Wilks' theorem does not apply. Even for the other models, we generally have too little data and too conservative a perspective to assume the asymptotic distribution, so we instead determine significance by comparing the observed likelihood ratio to the distribution of that statistic over the data with permuted vaccine-placebo labels. Here we show that, as expected, the power of this test for any chosen alternative is highest when data are drawn from that alternative model, and explore power for a realistic scenario.

For the simulation study, we generated data for a single site in a scenario similar to the RV144 sieve analysis: 44 vaccine-recipient subjects, 66 placebo-recipient subjects, with an expected 85% of the (null/placebo) values falling in the insert category, 12.5% in the second-most-frequent category, and the remaining 2.5% distributed evenly among the remaining 19 categories. This approximately matches the distribution among the nine “focus sites” used for the RV144 v1v2-focused sieve analysis. For all analyses, we use pseudocounts / hyperpriors  $\gamma$  which are even and sum to 1 (so with 21 categories, each is  $\alpha_{-1,i} = \frac{1}{21}$ ). For the “noninsert-monotonic” and “noninsert-unconstrained” models, we use the HIV-1 between-subjects substitution probability matrix (for 1% divergence) in both generating the data and in evaluating the likelihood. Since we are simulating with the insert category being 1, we are effectively restricting the values of  $\mathbf{p}_a$  to the first row of this substitution matrix, which corresponds to substitutions from Alanine to the other 19 amino acids, with 0 probability of transitioning to a gap category.

For each generated dataset, we evaluate posterior probabilities and p-values using each of the three alternative models ( $M_a^{io}$ ,  $M_a^{nm}$ , and  $M_a^{nu}$ ), both with and without using latent *sievable* indicators. When we use them, probabilities are set for half (22) of the simulated vaccinated subjects at  $p_{r,i} = 75\%$  and at  $p_{r,i} = 25\%$  for the other half. For each of these six models, we generated 1000 datasets from the specified alternative model at actual sieve

effect strengths of 0 (for size determination) and 30%. Latent indicators are also drawn for each dataset. Note that the datasets for the strength-0 analyses are the same for all model evaluations, whereas the datasets for the strength-30 analyses are the same among all evaluations that use latent *sievable* indicators, and the same among those that do not, but they necessarily differ between those two groups.

Each of the six models is evaluated using all four permutations of the following two binary options:  $\mathbf{p}_0$  can be treated as a “fixed effect” or as a “random effect”, and the flavor can be “one-phase”, with the placebo data included directly in the model, or it can be “two-phase”, with the placebo data used to estimate  $\mathbf{p}_0$  (or  $\boldsymbol{\alpha}_0$  when using “hierarchical”) and excluded from the likelihood calculation. In addition, for comparison we use Fisher’s exact test on the observed counts and also, separately, on the dichotomized counts (insert vs non-insert). We also compare each method to the non-parametric t-test from Gilbert et al. (2008).

[Figure 1 about here.]

Note that an exploration (data not shown) of two-phase models with various pseudocounts reveals sensitivity to their total: if the total is too small (or if pseudocounts are not used), then the data becomes extremely rare or even impossible and the power of the test suffers. Likewise if the total is large, the hyperprior dominates the data and the again the power suffers. This sensitivity is particularly acute when evaluating the approaches based on posterior probabilities.

#### 4. Discussion and Future Work

In this article we introduce a framework for a model-based approach to sieve analysis. The framework provides a useful platform for reasoning about sieve effects, and it is sufficiently general to accomodate simple dichotomizing models as well as relatively complex models that reflect different hypotheses about the nature of sieving. In particular, we introduce a Bayesian

and a Bayesian-frequentist hybrid approach to evaluating three new models of sieving, each of which comes in two “flavors”, depending on their treatment of the placebo-recipient data. One could argue that these models are each actually families of models, indexed by the parameters governing the prior distributions (and, for the “noninsert-monotonic” and “noninsert-unconstrained” model families, the value of  $\mathbf{p}_a$ ). The framework generally supports models of categorical sequence data in which sieve effects reduce the probability of the insert category and redistribute the removed mass among the remaining categories. The fundamental perspective is that of latent per-subject (and per-locus) indicators of being *sieved*. These indicators are modeled as independently  $\text{bernoulli}(p_s)$  distributed, with the sieve effect strength parameter  $p_s$  being the primary parameter of inferential interest.

The framework and models also support an additional per-subject (and per-locus) latent variable *sievable*, which when treated as a random variable effectively modifies the probability of being *sieved* by a per-subject multiplicative factor ( $p_{r,i}$ ). This *sievable* indicator framework allows for incorporation of arbitrary covariate models linked through each subject’s probability of being *sievable*. We have illustrated that if these probabilities are well-estimated, the power to estimate sieve effects is improved by their inclusion in the model. Our use of these probabilities has so far been restricted to the case in which they are deterministic and known, but with sufficient data it should be possible to simultaneously estimate parameters for a linear model of these probabilities as a function of covariates, though additional constraints would be required to ensure identifiability between these probabilities and  $p_s$ . We leave this to future work.

While the models explicitly address only a single site, multi-site analysis proceeds naturally. For testing multiple sites using multiple tests, we simply perform site-specific tests and apply a form of multiplicity adjustment. By setting the prior probability of the alternative model(s) at each site to a sufficiently low value, we accomplish a Bayesian form of multiplicity



adjustment. Such site-specific priors can also be used to accommodate prior evidence of sieve effects across sites. In a frequentist context, these priors are irrelevant, but standard multiplicity-adjustment methods may be applied across site-specific p-values as desired. For global testing (or for testing any set of multiple sites), posterior probabilities of the null and alternative models can simply be multiplied according to the logic of the inquiry. For instance, the posterior probability that “at least one” site is sieved is the complement of the probability that none are sieved, which is simply the product of the posterior probabilities of the null hypothesis. In a frequentist context, these probabilities can be compared to a null distribution derived by permuting treatment labels, as we do for individual sites.

The models that we introduce in this article are but a taste of the possible models that can be accommodated by this framework. One immediate criticism of these particular models is that they assume that sieving targets solely the insert category. The “noninsert-monotonic” model, for instance, can’t represent a sieve effect that lowers the probability of a non-insert category. If an antibody induced by the vaccine binds to any hydrophobic residue, then we might expect the sieve effect to lower the probability of all hydrophobic amino acids. A model reflecting this could easily be constructed using this framework (by applying the sieve effect strength multiplier  $1 - p_s$  to all hydrophobic amino acids, for instance), but we leave this too to future work.

We also leave to future work the extension of these methods to address multiple sequences per subject, which we envision could be accommodated fairly simply by representing each subject’s sequence-category as an unknown quantity with a multinomial distribution over the frequencies observed across the subject’s multiple sequences. This same approach could be used (instead, or in addition) to accommodate alignment uncertainty (via the posterior distribution of each amino acid at the particular locus, averaged over all possible alignments). More sophisticated approaches (to multiple observed sequences) are probably worth explor-

ing, though, such as extending the framework to represent all of a subjects’ sequences, with the subject-specific “sieved” indicator values shared across sequences.

Sieve analysis methods that condition on infected subjects are subject to the critique of post-randomization subgroup analysis, which hinders causal interpretability even in the context of a randomized controlled trial. To our knowledge, the only site-specific sieve analysis method that escapes this critique is the genotype-specific vaccine efficacy method (Prentice et al., 1978; Gilbert, 2000), which employs survival analysis approaches to include all of the clinical trial subjects in the analysis. This method necessitates dichotomization based on insert match/mismatch, so it can’t presently be used to address hypotheses about changes to the non-insert distribution, but there is clearly room to develop methods that can. At the least, it seems worth exploring incorporation of time-to-event information into the model-based sieve analysis framework (which could be done presently with reasonable assumptions about how time-to-event information can inform probabilities of *sievability*).

At present, no sieve analysis method can distinguish between *true sieve effects*, which affect the distribution of infecting sequences, and *post-acquisition effects*, which affect the distribution of later sequences through selective pressure on the evolving infection. The ideal solution is improved sampling strategies, which could potentially identify infections in a sufficiently-early phase such that post-acquisition changes, if present, could be identified. Short of that, developing analysis methods and statistical tests that are uniquely sensitive to acquisition effects or to post-acquisition effects could help distinguish among them. We have shown that the model-based framework supports tests that are sensitive to particular alternative hypotheses. It remains to be seen whether this power can be leveraged to distinguish true sieve effects from post-acquisition effects, but we argue that this work is a step in the right direction.

## 5. Estimating genotype-specific vaccine efficacy

Gilbert et al. (1998) discussed applying a multinomial logistic regression (MLR) model (?) for estimating the probabilities of the categories in a categorical sieve analysis. They showed that under certain assumptions, the parameters of this model are interpretable as log ratios of strain-specific relative risks of infection. Here we show that this result also applies to the models introduced in this paper, which are constrained special cases of the sieve MLR model. The sieve MLR model specifies that category-probabilities depend on vaccination status ( $z$ ) according to

$$P(Y_j = y|z) = \frac{\exp(\beta_{0,y} + z\beta_{1,y})}{1 + \sum_{i=2}^k \exp(\beta_{0,i} + z\beta_{1,i})}, \quad (10)$$

where  $\beta_{0,1}$  and  $\beta_{1,1}$  are 0 to ensure identifiability. This can be seen as a reparameterization of a completely unconstrained model with arbitrary category probabilities for placebo recipients given by

$$p_{0,y} = P(Y_j = y|z = 0) = \frac{\exp(\beta_{0,y})}{1 + \sum_{i=2}^k \exp(\beta_{0,i})},$$

and arbitrary category probabilities for vaccine recipients given by

$$p_{1,y} = P(Y_j = y|z = 1) = \frac{\exp(\beta_{0,y} + \beta_{1,y})}{1 + \sum_{i=2}^k \exp(\beta_{0,i} + \beta_{1,i})}.$$

Gilbert et al. (1998) discussed the ordered stereotype model (?) as a variant of the MSA model in which, to enforce monotonicity across the ordered categories, additional constraints are enforced on the parameters of (10). They go on to prove that when assuming a proportional hazards model for the overall (marginal) relative risk of infection, and when additionally assuming a proportional hazards model for category-specific per-exposure-event probabilities of infection, the parameters  $\beta_1$  of both the MSA model and the ordered stereotype model are not only interpretable as log-ratios of *retrospective* category-specific relative risks which condition on both exposure and infection, but are also interpretable as log-ratios of *prospective* relative risks of exposure-and-infection-by viruses in each particular category. They show that this result holds for the MSA model (with arbitrary parameter

constraints) so long as three additional assumptions are met: multiple infection is not possible during the trial, the relative conditional probability of exposure to each strain is constant over time during the trial, and the exposure distribution is the same across treatment arms.

#### ACKNOWLEDGEMENTS

The author thanks Peter Gilbert for helpful guidance and critical reading of early versions of the paper, and the rest of the SCHARP sieve analysis group in the Vaccine and Infectious Disease Division of the Fred Hutchinson Cancer Research Center for helpful feedback and inspiration, especially Craig A. Magaret, Tomer Hertz, Allen deCamp, and (external member) Morgane Rolland. The author is deeply grateful also to the DAIDS-funded HIV Vaccine Trials Network and to the NIH, which each provided support for this project, and to the participants in HIV-1 vaccine trials, whose action gives hope.

#### REFERENCES

- Berman, P., Gray, A., Wrin, T., Vennari, J., Eastman, D., Nakamura, G., et al. (1997). Genetic and immunologic characterization of viruses infecting MN-rgp120-vaccinated volunteers. *Journal of Infectious Diseases* **176**, 384–397.
- Gilbert, P. (2000). Comparison of competing risks failure time methods and time-independent methods for assessing strain variations in vaccine protection. *Statistics in medicine* **19**, 3065–3086.
- Gilbert, P., Self, S., and Ashby, M. (1998). Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* **54**, 799–814.
- Gilbert, P., Self, S., Rao, M., Naficy, A., and Clemens, J. (2001). Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *Journal of clinical epidemiology* **54**, 68–85.

- Gilbert, P., Wu, C., and Jobes, D. (2008). Genome scanning tests for comparing amino acid sequences between groups. *Biometrics* **64**, 198–207.
- Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., J., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, pp. 541–554.

**Figure 1.** Powers of the various methods for each of three data-generation methods, depicted in brackets ([IO]: insert-only; [NM]: noninsert-monotonic; [NU]: noninsert-unconstrained), evaluated by Fisher’s exact tests, the GWJ t-test, and corresponding MBS model (eg the [IO]-generated data is evaluated using the insert-only model). The powers range from near 0 (white) to 100% (red). The “random  $p_0$ ” methods are not implemented for the NM and NU models, so are omitted. The IO and NU data are well-modeled by the dichotomized Fisher and the full Fisher models, respectively. The NC data are best modeled using MBS-NC, and in all cases the two-phase approach is more powerful than the one-phase approach.